

Title:**Joint Selection of Wavenumber Regions for MidIR and Raman Spectra and Variables in PLS Regression using Genetic Algorithms****Authors & affiliations:**

Lidwine Grosmaire¹, Pedro Maldonado¹, Christelle Reynès², Robert Sabatier², Dominique Dufour³, Thierry Tran⁴ and Jean-Louis Delarbre¹

¹ Laboratoire de Physique Moléculaire et Structurale – UMR Qualisud – Université Montpellier 1 - France

² Laboratoire de Physique Industrielle et Traitement de l'Information – EA 2415 - Université Montpellier 1 – France

³ CIRAD – Dpt Persyst - UMR Qualisud – CIAT, Cali, Colombia

⁴ CIRAD - Dpt Persyst - UMR Qualisud - CSTRU - Kasetsart University – Bangkok - Thailand

Abstract: (Your abstract must use **Normal style** and must fit in this box. Your abstract should be no longer than 300 words. The box will 'expand' over 2 pages as you add text/diagrams into it.)

This work fits into the context of cassava processing. Production and consumption of this product is steadily increasing worldwide and especially in tropical regions where, after harvest, cassava starch is extracted according to an empirical process: natural fermentation and sun-drying, which gives to this product an interesting breadmaking capacity despite of the absence of gluten.

The objective of this work is to try to explain the breadmaking ability from different parameters (physicochemical and spectroscopic data) using a statistical regression method while selecting variables of different types: individual and intervals.

In chemometrics, the choice of explanatory variables is a problem often discussed, but, when it comes to select intervals, methodologies are rarer and more complex (Höskuldsson 2001). Among the specific methods developed (Norgaard et al 2004), Genetic Algorithms (GA) were chosen and combined with the PLS method to select intervals (Leardi 2000, Leardi and Norgaard 2004).

In this case, the explanatory variables are organized in a multitable in which intervals and individual variables are selected in order to predict one variable of interest: the breadmaking capacity. To this end, we will use and adapt a GA developed in a context of discrimination (usual LDA), jointly with the PLS1 method (Reynes et al 2006), this method is called AGvPLSm.

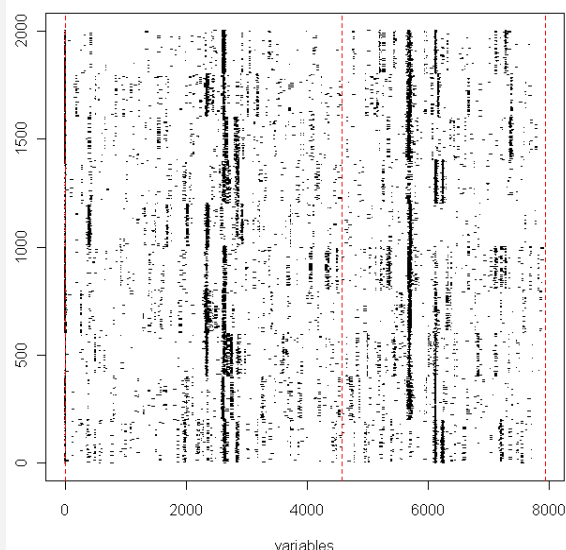


Figure 1 : Final GA populations characteristics: selected variables are indicated by black points

The variables selected by this method are shown in Figure 1 which confirms the global convergence of algorithms. AGvPLSm compared with other methods (cf table 1) shows that the resulting model is more efficient ($r^2 \approx 99\%$) and leads to a reasonable number of selected variables in the multitable (4% of the initial variables).

Method	# variables	# components	R ²	R ² _{CV}

PLS	7926	7	0.7836	0.6605
PLS + VIP	4	3	0.7210	0.6650
AGvPLSm	311	12	0.9936	0.8273

Table 1 : Comparison of different method results (number of selected variables, number of retained PLS components, R^2 and cross-validation R^2).

In terms of interpretation, this method allowed to highlight the importance of some physicochemical variables and select a small number of spectral intervals that significantly complete the model.

Höskuldsson, A. (2001) Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory System*, 55, 23-38.

Leardi, R. (2001) Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics*, 15, 559-569.

Leardi, R. Norgaard, L. (2004) Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Chemometrics and Intelligent Laboratory System*, 18, 486-497.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B. (2000) Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54, 3, 413-419.